CYCLADES IST-2000-25456
An Open Collaborative Virtual Archive Environment

Co-ordination Report Preliminary Version (D8.1.1)

_____

Abstract: This report is about the co-ordination activity among the CYCLADES project and
the Open Archive Initiative (OAI).

Introduction:

The proposed open archives environment aimed by Cyclades, consists of two components: the archives and the services. The former will be developed by the Open Archives initiative (OAI). This US initiative aims at guaranteeing interoperability among e-print archives. It has established a set of relatively simple but potentially quite powerful interoperability specifications that facilitate the development of services implemented by third parties.

CYCLADES will base the development of the service environment on these specifications. In particular, a core set of cross-archive value-added services will be developed to constitute a federation of independent but interoperable services. According to this approach, a service provides a functionality and can either work independently or can communicate and collaborate with other services to offer a new value-added service. The Service Environment will provide OAI compliant functionality.

Therefore a strong and constructive cooperation between Cyclades and the OAI is necessary to ensure the cohesion and adequacy of the archives and the services.

From the architectural point of view, OAI has defined a pragmatic, incremental and collaborative approach towards interoperability of e-print archives. This approach guarantees a minimal but potentially highly functional level of interoperability among archives. Such interoperability is beneficial to the archive and service providers.

CYCLADES will develop the service environment following a federative approach. According to this approach, a federation of autonomous but interoperable services through well defined communication protocols will be developed. These services, accessible via Internet, are visible to the users/scholars and accessible by them both separately or as a member of a cluster of interoperable services. This architectural approach gives the system great *flexibility*, permits system *clustering*, i.e., the aggregation of system services into subsets which provide more sophisticated services and allows an easy *customisation* of the system in order to be able to serve several and different communities.

Overview of the OAI conventions:

Within the OAI conventions, *archive* it is used to denote a component supporting the following set of functions:

- a submission mechanism
- a long term storage system
- a management policy for the submission of documents and their preservation
- an open machine interface, that enables third parties to collect data from the archive

In this framework, the notion of a *record* in an archive is used. Some archives may store metadata that describes full content without storing the full content itself. In this case, the metadata is a *record*. Other archives may also store full content. It is assumed that if full content is stored, there will always be associated metadata stored in the archive as well as a mechanism to tie metadata and content together. In this case the combination of metadata and full content is a *record*.

E-print archive interoperability is established using three mechanisms*:*

- The definition of a set of simple metadata elements – the Dublin Core Metadata Set (DCMS)- for the sole purpose of enabling coarse granularity document discovery among archives;

- The definition of a common protocol – the Open Archives initiative Protocol – to enable extraction of DCMS and archive-specific metadata from participating archives

- The agreement to use a common syntax, XML, to represent and transport both the DCMS and archive-specific metadata sets.

These mechanisms treat documents as black-boxes; archives can have idiosyncratic document representations. The mechanisms only specify a URL entry point to the archives' individual document models. The question and functionality of common services are left to implementers who wish to exploit these mechanisms in order to implement services on top of them.

Dublin Core Metadata Set: a collection of metadata elements intended to facilitate coarse granularity resource discovery among the records in distributed and dissimilar archives. The semantics of this set have purposely been kept simple in the interest of easy creation and widest applicability. There is no provision for qualification or extension of the nine elements. The expectation is that individual archives will maintain metadata with more expressive semantics and the Open Archives Initiative Protocol will provide the mechanism for retrieval of this richer metadata.

Open Archives Initiative Protocol: a set of protocol requests that are delivered via HTTP. This protocol is a subset of the full Dienst protocol. The protocol requests in the subset provide the following functionality:
- List the full identifiers for records stored in an archive. An optional argument permits the client to specify that the list should only include records added after a specific date. Another optional argument allows the client to specify that the records should be accompanied by the metadata associated with the identifier.
- Return the metadata for a specific record in a requested format.
- Return the list of metadata formats supported by an archive.
- Return the list of metadata formats available for a specific record.
- Return the structure of the partitions by which an archive is organised.
All responses to these requests are formatted in XML.

Organisational Aspects of the Open Archives Infrastructure

An organisational framework has also been defined in order to facilitate the implementation of the technical components as well as to establish a communication mechanism between data providers and service providers.

For the data providers, the framework contains some recommendations related to the implementation of the previously described technical components of the infrastructure. In addition, the following important organisational elements have been introduced:
- The notion of a *unique archive* identifier to unambiguously represent each e-print archive, as well as the notion of *full identifier* of a record in an archive. Since *this full identifier* is a concatenation of the unique archive identifier and the unique persistent identifier of a record in an archive, this *full identifier* will be persistent.
- The recommendation to document metadata formats other than the DCMS, as well as a facility to share this documentation with data and service providers.
- A facility to register an e-print archive as being compliant with the Open Archives Infrastructure, by means of the provision of a filled-out version of a data provider template that describes crucial characteristics of the archive.
- A list of e-print archives that comply with the Open Archives Infrastructure, from which links are available to the documents describing their crucial characteristics.

For the service providers, the following important organisational elements have been introduced:

- The request to maintain the original full identifiers of the records harvested by the service provider.
- The request to comply with the terms and conditions that data providers have brought forward in their filled-out data provider template.
- A facility to register a service as being compliant with the Open Archives Infrastructure, by means of the provision of a filled-out version of a service provider template that describes aspects of the service.
- A list of services that comply with the Open Archives Infrastructure, from which links are available to the filled-out templates that describe them.

As a token of the fianl validation, several archives have currently expressed their intention to adopt the OAI agreements (http://www.openarchives.org/).

Having these elements in mind, a co-ordination of the validation and experimentation activities of the CYCLADES project with OAI appears highly necessary. The validation and experimentation results will be made available to OAI for undertaking the necessary corrective actions.

Regular working meetings between representatives of the two projects will be planned. A workshop to be promoted by the OAI Technical Committee will be organised in cooperation with CYCLADES next September in co-ordination with the ECDL 2001 conference.

Project Co-ordination tasks

In that perspective, Work Package 8 was included in the work-programme. It consists in continuously monitoring the progress achieved by CYCLADES and OAI. This monitoring should guarantee that the interoperability agreements defined by OAI and their evolution is correctly understood and used by the service environment to be developed by CYCLADES.

It has also to guarantee that the impact on the service level of new communication protocols and standards (metadata evolution)or changes to interoperability agreements and technical recommendations introduced at the archive level (and vice versa) is thoroughly evaluated and feedback is sent to the originators of the changes.

Monitoring and Interoperability Verification

A continuous monitoring activity of the OAI (meetings and related papers)is in progress in order to guarantee that the assumptions under which the Cyclades services work still hold. Both the old protocol version 1.0 and current protocol version 1.1 have been read and a consistency of Cyclades with regards to .this protocol has been checked out. Interoperability is till guaranteed.Version 1.1 does now comply with the W3C XML Schema Recommendation of May 2001.

On top of this, concerning compatibility, while the Cyclades system will use the protocol specified by the Open Archives Initiative (OAI) to harvest metadata from any number of archives that support the OAI standard, the harvesting will be done by one of several interoperable services, Cyclades is not restricted to using open archives, as additional services can be added later, supporting other kinds of electronic archives (See deliverable D2.1.1, Global System Architecture Report).

Subscription to the OAI mailing lists has been performed.

Furthermore, the list of registered repositories as well as the register OAI services is continuously monitored.

The present registered repositories and actual service have been identified and are the following:

To date, July 25,2001,**the list of registered repositories** is:

OAI Repository Identifier	Repository Name
celebration	A Celebration of Women Writers
anlc	Alaska Native Language Center
aps	American Philosophical Society
arXiv	arXiv
bmc	BioMed Central
CDLCIAS	California International and Area Studies Digital Repository
caltechCSTR	Caltech Computer Science Technical Reports
cimi	CIMI Metadata Harvesting Working Group Demonstration Repository
citebase	Cite-Base services
cogprints	CogPrints
Comparative	Bantu Online Dictionary (CBOLD)
CDLDERM	Dermatology Digital Repository
elra	European Language Resources Association
formations	Formations
cav2001	Fourth International Symposium on Cavitation
HUBerlin	Humboldt University of Berlin, GERMANY, Document Server
scout	Internet Scout Project OAI Repository
lcoa1	Library of Congress Open Archive Initiative Repository 1
ldc	Linguistic Data Consortium
	LTRS
	M.I.T. Theses
	NACA
NSDL-DEV-CU	NSDL Open Archives Server at Cornell University
OCLC	Online Computer Library Center Theses and Dissertations Repository
	Perseus Digital Library
physdoc	PhysNet, Oldenburg, Germany, Document Server
	Resource Discovery Network
sceti	Schoenberg Center for Electronic Text and Image
lacito	The LACITO Archive
dfki	The Natural Language Software Registry
	The Oxford Text Archive
CDLTC	Tobacco Control Digital Repository
DUETT	Tropicos
tkn	University Library, Gerhard-Mercator-Universitaet Duisburg
	University of Tennessee Libraries

To date, July 25,2001,the **list of service** is:

- **Arc** (service) Old Dominion University. A federated search services based on metadata harvested from several OAI compliant repositories.

- **Repository Explorer** (service) Virginia Tech. An interactive, web-based tool to test repositories for compliancy with the Open Archives Initiative Protocol version 1.0.

- **TORII** (service, template) International School for Advanced Studies, Trieste, Italy. Unified access to various open archives (Physics and Computer Science).Filtering and advanced searching. Personalization. For more information, see the TIPS consortium web pages at http://tips.sissa.it.

Additionally, a first metadata gathering activity has been started to determine some statistics about the to be gathered data from the OAI compliant archives. Some non-uniformity of the data has been detected. The report of this gathering process is submitted to a OAI Workshop and maybe considered as a feedback activity. The gathered data seems already to be enough for making the system appealing.

Meetings

The Cyclades project has been presented at the OAI, Open Meeting in Berlin, Germany, February 26,2001,where the meeting marked the European public release of the specifications of the OAI interoperability architecture. The release represents the fixing of these specifications for a one year period of experimentation. The meeting included an overview of the organisational and historical context of the Open Archives Initiative, a detailed presentation of the OAI interoperability architecture, presentations by implementers who have worked with or plan to work with protocol, and opportunities to discuss building of communities within the OAI technical framework. A U.S. companion to this meeting was held in Washington, DC on January 23,2001.

A first meeting has been held in July 2001 with Carl Lagoze, the representative of Cornell University and heavy involved in the OAI, discussing about the project that finally has been approved by NSF. We expect to closely collaborate and a human resource exchange program is under the study.